Efficient RGB-T Tracking via Cross-Modality Distillation

Tianlu Zhang[1] Hongyuan Guo[1] Qiang Jiao[1] Qiang Zhang[1*] Jungong Han[2,3]
[1]School of Mechano-Electronic Engineering, Xidian University, China
[2]Department of Computer Science, the University of Sheffield, UK.
[3]Centre for Machine Intelligence, the University of Sheffield, UK.
{tianluzhang, hyg}@stu.xidian.edu.cn, {qzhang, qjiao}@xidian.edu.cn, jungonghan77@gmail.com

Abstract

Most current RGB-T trackers adopt a two-stream structure to extract unimodal RGB and thermal features and complex fusion strategies to achieve multi-modal feature fusion, which require a huge number of parameters, thus hindering their real-life applications. On the other hand, a compact RGB-T tracker may be computationally efficient but encounter non-negligible performance degradation, due to the weakening of feature representation ability. To remedy this situation, a cross-modality distillation framework is presented to bridge the performance gap between a compact tracker and a powerful tracker. Specifically, a specific-common feature distillation module is proposed to transform the modality-common information as well as the modality-specific information from a deeper two-stream network to a shallower single-stream network. In addition, a multi-path selection distillation module is proposed to instruct a simple fusion module to learn more accurate multi-modal information from a well-designed fusion mechanism by using multiple paths. We validate the effectiveness of our method with extensive experiments on three RGB-T benchmarks, which achieves state-of-the-art performance but consumes much less computational resources.

1. Introduction

RGB-T tracking is the task of estimating the state of an arbitrary target in each frame of an RGB-T video sequence [35]. Due to the affordability of thermal infrared (TIR) sensors, RGB-T tracking draws more and more research interest.

As shown in Fig. 1 (a), most existing RGB-T tracking models first adopt a two-stream structure to extract multi-level unimodal RGB and TIR features, respectively, and then employ elaborate-designed multi-modal feature fusion modules to exploit complementary information within the multi-modal data. Finally, they deduce the target state, often represented by a bounding box, from the fused features. Although great progress has been made, these powerful RGB-T tracking models usually require high computational costs and large model sizes to handle the information of two modalities in the stages of unimodal feature extraction and multi-modal feature fusion.

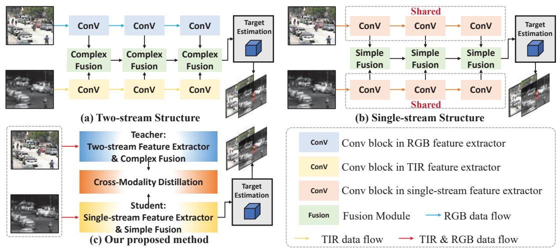

Figure 1. Architectures of different RGB-T tracking models. (a) Two-stream structure. (b) Single-stream structure. (c) Our proposed method.

There are two straightforward solutions to tackle the complexity and efficiency issues. One is to adopt a single-stream feature extractor with fewer convolutional layers, and the other is to employ simpler multi-modal feature fusion modules, as shown in Fig. 1 (b). Although such compact models can reduce computational complexity, they inevitably bring non-negligible performance degradation due to the weakening of unimodal feature representation ability and multi-modal complementary information exploration ability. For instance, a powerful RGB-T tracker [35] with a two-stream structure and complicated multi-modal feature fusion modules suffers from severe performance degradation after the above model simplification operations (84.4% precision rate vs 78.1% precision rate on RGBT234 dataset [11]), as shown in Fig. 2.

Now, the research question becomes: can we shrink the RGB-T tracker without sacrificing performance? This paper answers this question using knowledge distillation, which allows a compact model to obtain a similar ability of a complex model at little cost. We call this complex but powerful model the teacher model, and call this

*Corresponding author.

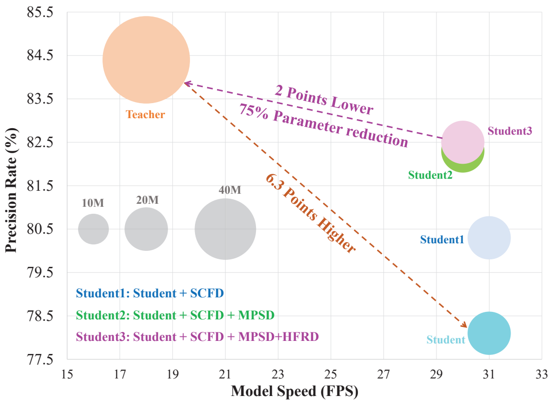

Figure 2. Experimental results of different RGB-T tracking structures on RGBT234 dataset [11]. Teacher denotes the two-stream structure with complex fusion modules. Student denotes a single-stream structure with simple fusion operations. The teacher model employs ResNet50 [7] for feature extraction and fusion modules in [35] for multi-modal feature fusion, respectively. The student model employs ResNet18 [7] for feature extraction and concatenation for multi-modal feature fusion, respectively.

compact model the student model. Although some works [15, 19, 20, 29] have made considerable progress on knowledge distillation in multi-modal tasks, they fail to conduct a deep investigation on the huge feature differences between teacher and student in the unimodal feature extraction stage as well as in the multi-modal feature fusion stage, thereby resulting in suboptimal efficiency of the knowledge transformation. For that, a novel teacher-student knowledge distillation training framework, named Cross-Modality Distillation (CMD), is proposed to elaborately guide efficient imitation from three stages: unimodal feature extraction, multi-modal feature fusion and target estimate estimation, as shown in Fig. 1 (c).

Specifically, in the stage of unimodal feature extraction, as pointed out by many previous works [9, 17], the shallower layers of unimodal features usually contain abundant low-level spatial details, which are usually modality-dependent. Differently, the deeper layers of unimodal features often contain many high-level semantic cues, which tend to be strongly modality-consistent. The student model uses a compact single-stream network to extract both RGB features and TIR features, which not only lacks the ability to extract modality-specific information in the shallower layers, but also insufficiently explores the modality-common information in the deeper layers. These interesting observations inspire us to design a Specific-common Feature Distillation (SCFD) module, which transforms the modality-specific information as well as the modality-common information from a two-stream deeper network to a single-stream shallower network.

Second, in the stage of multi-modal feature fusion, the complex multi-modal feature fusion modules in the teacher model show great advantages in various scenarios, while the simple fusion strategies in the student model are usually effective in some specific scenarios. It is difficult for a student model with a single simple fusion strategy to learn more effective complementary information mining capabilities from a complex teacher model due to the large feature differences. Therefore, we design a fusion module with multiple simple fusion strategies in the student model, denoted as Multi-path Selection Distillation (MPSD) module. In the process of learning from the teacher model, the student model can adaptively combine different types of fusion features to make up for the lack of complementary information mining capabilities of a single simple fusion strategy.

Finally, in the stage of target state estimation, with the weakening of the feature representation ability of the student model, the discriminative ability of the tracker for distractors is also reduced. For that, we further present a Hard-focused Response Distillation (HFRD) module to improve the student model's discriminative ability by alleviating the problem of data imbalance between the targets and the backgrounds, which employs the response maps generated by the teacher model to instruct the student to focus on distinguishing targets from hard negative samples.

As shown in Fig. 2, each of our proposed modules continuously reduces the performance gap between the student model and the teacher model without increasing the number of parameters obviously. To sum up, our work improves an RGB-T tracker dramatically because of the following two contributions:

- A Cross-Modality Distillation (CMD) framework is presented to bridge the performance gap between a compact student model and a powerful teacher model through three stages, i.e, unimodal feature extraction, multi-modal feature fusion and target state estimation. To the best of our knowledge, we are the first to introduce knowledge distillation for multi-modal tracking.
- Experimental results show that our proposed approach helps a student model achieves the state-of-the-art performance on the challenging GTOT [10], RGBT234 [11] and LasHer [14], while reducing the number of parameters and computational complexity.

2. Related Work

RGB-T Tracking Methods. The past few years have witnessed the increase of RGB-T tracking algorithms [4, 10, 13, 17, 31, 39]. Among them, numerous RGB-T trackers [6, 17, 30, 39] have been presented based on the MDNet [18]. For instance, in [17], Li *et al.* introduced a multi-adapter architecture to learn modality-common, modality-specific and instance-aware target representations, respectively. In [39], Zhu *et al.* first presented a network to aggregate the features from all of the layers and all of the modalities. After that, these aggregated features were further pruned to reduce noise and redundancy. Recently, Zhang *et al.* [35] introduced DiMP [1] as their baseline tracker and achieved promising tracking performance. Meanwhile, aiming to speed up the tracking, some works [36, 37] bring the Siamese networks to RGB-T tracking, where their models are trained in an offline manner. Although some progress has been made, these models still suffer from the limitations of large model sizes and high computation costs.

Knowledge Distillation Methods. Knowledge Distillation (KD) was first proposed by Hinton *et al.* [8] to pass dark knowledge from complicated teachers to compact students, enabling students to maintain strong performance as teachers. FitNet [21] proves that the semantic information from intermediate features is also helpful to guide the student model. Besides image classification, KD is widely applied to object detection and object tracking tasks. For instance, [2] and [16] used KD to speed up the detection and segmentation networks, respectively. Furthermore, Wang [24] proposed to learn a more compact backbone for faster feature extraction in correlation filter based trackers. Shen *et al.* [22] used KD to compress deep Siamese-based trackers for high-performance visual tracking.

In addition, KD is also employed for some multi-modal tasks, such as RGB-D salient object detection [19, 20] and RGB-T pedestrian detection [15, 29]. Specifically, in [20, 29], the single-stream feature extractors and the early fusion strategies were employed in their student models. However, both of them only simply employ the distillation loss functions to improve the performance of student models by using fused features or label knowledges of teacher models, and pay less attention to the huge differences between the teacher model and the student model in the unimodal feature extraction stage as well as in the multi-modal feature fusion stage. Differently, in this paper, we aim to narrow the feature differences between the student model and the teacher model by specifically learning strategies at multiple stages.

3. Distilled RGB-T Tracking

Given a powerful teacher model for RGB-T tracking, the proposed CMD framework aims to prompt a more efficient student model to learn from the teacher model. The knowledge from the teacher model is transferred to the student model to mimic more effective feature representation. This section starts with an overview of the proposed CMD framework. Then, we briefly provide an introduction of the employed teacher and student models. Finally, the three proposed knowledge distillation modules (i.e., SCFD, MPSD and HFRD) are described in details.

3.1. Overview

As illustrated in Fig. 3, the proposed CMD framework includes a teacher model, a student model and three knowledge distillation modules. The teacher model takes a pair of RGB-T images as input, and employs a two-stream feature extractor and several complex multi-modal feature fusion modules for unimodal feature extraction and multi-modal fusion, respectively. Finally, the fused features will be fed into the target state estimation module to obtain the final tracking results. Different from the teacher model, the student model uses a single-stream feature extractor and several efficient multi-modal fusion modules. Although the student model has a higher running speed, the simplification of the model inevitably leads to a decrease in tracking performance.

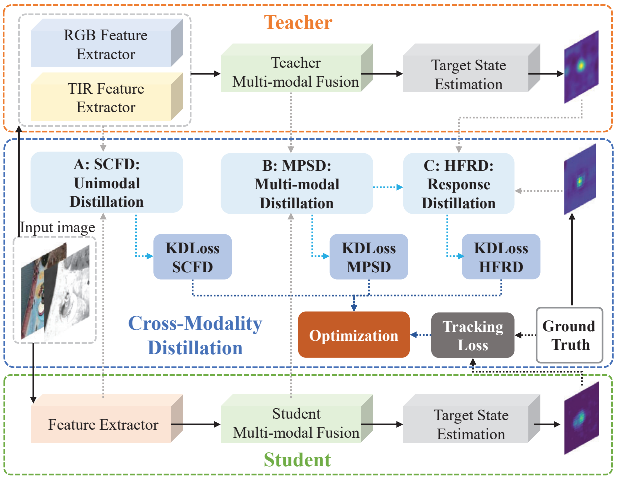

Figure 3. Overview of the proposed CMD framework.

To make up for the huge performance gap between the student model and the teacher model, the proposed CMD framework attempts to coach the learning process of the student model from three stages: unimodal feature extraction, multi-modal feature fusion and target state estimation. Accordingly, in the first stage, by using a proposed SCFD module, the powerful two-stream feature extraction network of the teacher model will transfer such modality-specific information as well as modality-common information into the single-stream network of the student model to enhance its representation ability for unimodal features. In the second stage, we will present an MPSD module to shrink the differences between the fused features obtained by the teacher model and those obtained by the student model via a multi-path optimization strategy. In the third stage, with a proposed HFRD module, we will adopt the response map generated by the teacher model in a form of spatial attention to instruct the student model to focus on the discrimination of difficult samples, thereby improving its discrimination ability. The improvements in the above three stages will effectively narrow the performance gap between the student model and the teacher model, enabling the student model to achieve competitive tracking results with the teacher model but with fewer parameters and higher computational efficiency.

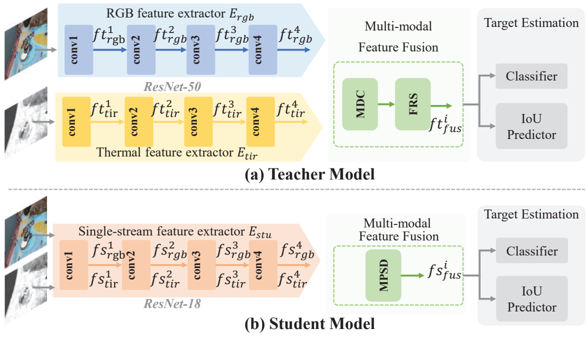

Figure 4. Structure of the employed teacher model and the student model.

3.2. Teacher and Student Model

In this section, we will describe the architectures of the employed teacher and student models, which are both based on the recent deep RGB tracker DiMP [1]. As shown in Fig. 4, both the teacher model and the student model can be divided into three stages: unimodal feature extraction, multi-modal feature fusion and target state estimation.

Feature extraction. In the teacher model, two feature extractors, denoted as E_{rgb} and E_{the}, regard RGB and TIR modalities in parallel. The two feature extractors both adopt ResNet50 [7] as the backbone to extract multi-level RGB and TIR features, as shown in Fig. 4 (a). Differently, in the student model, only one feature extractor, denoted as E_{stu}, regards both RGB and TIR modalities simultaneously. As shown in Fig. 4 (b), E_{stu} just adopts ResNet18 [7] as the backbone for simplification. Similar to the original DiMP tracker, in both the teacher model and the student model, we use the features from block3 and block 4 for regression, and those features only from block4 for classification. The extracted RGB and TIR features from the teacher model are denoted as ft^i_{rgb} and ft^i_{tir}, respectively, and the extracted RGB and TIR features from the student model are denoted as fs^i_{rgb} and fs^i_{tir}, respectively, where $i \in \{1, 2, 3, 4\}$ indexes the feature level.

Multi-modal feature fusion. By performing the multi-modal fusion modules on the $3rd$ and $4th$ levels of RGB and TIR features, we obtain the fused features ft^3_{fus} and ft^4_{fus} in the teacher model and the fused features fs^3_{fus} and fs^4_{fus} in the student model, respectively. Our teacher model employs a Modality Difference Compensation (MDC) module and a Feature Re-selection module (FRS) for multi-modal feature fusion [35]. Differently, our student model utilizes the proposed MPSD modules for multi-modal feature fusion. Details of MPSD will be introduced in Section 3.4.

Classification and regression. Finally, these fused features will be fed to the classification and regression heads, which have the same architectures with those in the original DiMP. Especially, in this stage, the student and teacher models both apply the original classification and regression heads in DiMP. We refer readers to [1, 5] for more details.

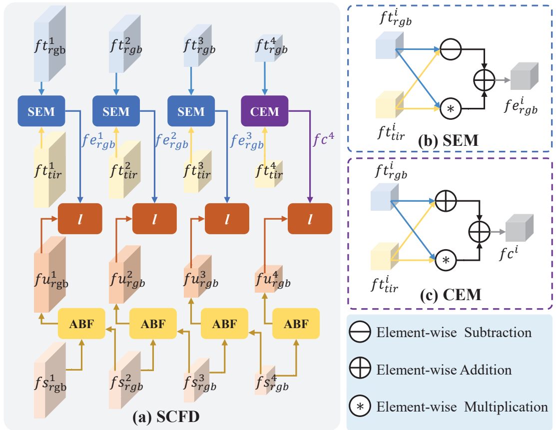

Figure 5. The proposed specific-common feature distillation module for the RGB modality. The SCFD module for the TIR modality employs the same architecture as that in the RGB modality.

3.3. Specific-Common Feature Distillation

This section elaborates on the proposed SCFD module for the two-stage unimodal feature distillation, which lets the single-stream feature extraction module in the student model enable to learn the modality-common information as well as the modality-specific information from the teacher model, as shown in Fig. 5 (a).

We first perform cross-modal interaction on the unimodal RGB features and TIR features from the teacher model to highlight the modality-common information and modality-specific information at different layers, respectively, for better guiding the learning of the student model. Specifically, as shown in Fig. 5 (b), given the unimodal features of shallow layers (i.e., $\{ft^i_{rgb}|i = 1, 2, 3\}$ and $\{ft^i_{tir}|i = 1, 2, 3\}$) from the teacher model, the proposed Specific Enhanced Modules (SEMs) are employed to obtain such modality-interacted features fe^i_{rgb} and fe^i_{tir} ($i = 1, 2, 3$) with more modality-specific information via subtraction and multiplication. Mathematically,

$$\begin{aligned} fe^i_{rgb} &= (ft^i_{rgb} \circledast ft^i_{tir}) \oplus (ft^i_{rgb} \ominus ft^i_{tir}), \\ fe^i_{tir} &= (ft^i_{rgb} \circledast ft^i_{tir}) \oplus (ft^i_{tir} \ominus ft^i_{rgb}), \end{aligned} \tag{1}$$

where $\ominus$, $\oplus$ and $\circledast$ denote element-wise subtraction, element-wise addition and element-wise multiplication, respectively. $ft^i_{rgb} \circledast ft^i_{tir}$ reflects the jointly valid information within RGB and TIR features. While, $ft^i_{rgb} \ominus ft^i_{tir}$ represents the modality-specific information of the RGB modality with respect to the TIR modality. Similarly, the modality-specific information of the TIR modality with respect to the RGB modality can be obtained by $ft^i_{tir} \ominus ft^i_{rgb}$. Accordingly, fe^i_{rgb} and fe^i_{tir} highlight such modality-specific information in addition to preserve jointly valid information, which can be applied to guide feature learning of the student model in shallow layers.

Alternatively, for the RGB and TIR features of deep layers (i.e., ft^4_{rgb} and ft^4_{tir}), the proposed Consistence Enhanced Module (CEM) is employed to obtain modality-interacted features fc^4 with more modality-common information via addition and multiplication, as shown in Fig. 5 (c). Mathematically,

$$fc^4 = (ft^4_{rgb} \circledast ft^4_{tir}) \oplus (ft^4_{rgb} \oplus ft^4_{tir}). \tag{2}$$

Here, by applying the element-wise addition on ft^4_{rgb} and ft^4_{tir}, the consistency of high-level semantic cues within multi-modal data can be further enhanced. Therefore, fc^4 can better guide the learning of the student model in deep layer.

With the modality-interacted features from the teacher model, the next step is to adjust the feature-channel dimensions of the student model to be consistent with those of the teacher model. Here, inspired by the idea of Knowledge Review [3], we employ a series of attention based fusion (ABF) modules [3] to adjust the channel dimensions of unimodal features and dynamically aggregate the cross-layer features in the student model. The modified features of the student model from ABFs (i.e., $\{fu^i_{rgb}|i = 1,2,3,4\}$ and $\{fu^i_{tir}|i = 1,2,3,4\}$) and the modality-interacted features of the teacher model (i.e., $\{fe^1_{rgb}, fe^2_{rgb}, fe^3_{rgb}, fc^4\}$ and $\{fe^1_{tir}, fe^2_{tir}, fe^3_{tir}, fc^4\}$) will be employed together to force the student model to mimic the specific and common information from the teacher model via a proposed feature-learning distillation loss L_{SCFD}, which is formulated as:

$$\begin{aligned} L_{spe} &= \textstyle\sum_{i=1}^{3} l\left(fe^i_{rgb}, fu^i_{rgb}\right) + \sum_{i=1}^{3} l\left(fe^i_{tir}, fu^i_{tir}\right), \\ L_{com} &= l\left(fc^4, fu^4_{rgb}\right) + l\left(fc^4, fu^4_{tir}\right), \\ L_{SCFD} &= L_{spe} + L_{com}, \end{aligned} \tag{3}$$

where $l(*)$ denotes the standard MSE loss used in [21].

3.4. Multi-path Selection Distillation

In order to learn the exploration ability of complementary information from the teacher model more effectively, we design a fusion module by using multiple fusion strategies, denoted as Multi-path Selection Distillation (MPSD) module, in the student model. In the process of learning from the teacher model, the student model can adaptively optimize the pathes to reduce feature differences.

Specifically, in the student model, the proposed MPSD module first performs multi-modal feature fusion from three typical perspectives: modality differences, modality commonality and modality complementary. Given the original RGB features fs^i_{rgb} and TIR features fs^i_{tir} from the $3rd$ and $4th$ levels in the student model, three types of initially fused features $fs^i_{fus,1}$, $fs^i_{fus,2}$ and $fs^i_{fus,3}$ are computed by

$$\begin{aligned} fs^i_{fus,1} &= sa(fs^i_{rgb}, fs^i_{tir}), \\ fs^i_{fus,2} &= fs^i_{rgb} \circledast fs^i_{tir}, \\ fs^i_{fus,3} &= fs^i_{rgb} \ominus fs^i_{tir}. \end{aligned} \tag{4}$$

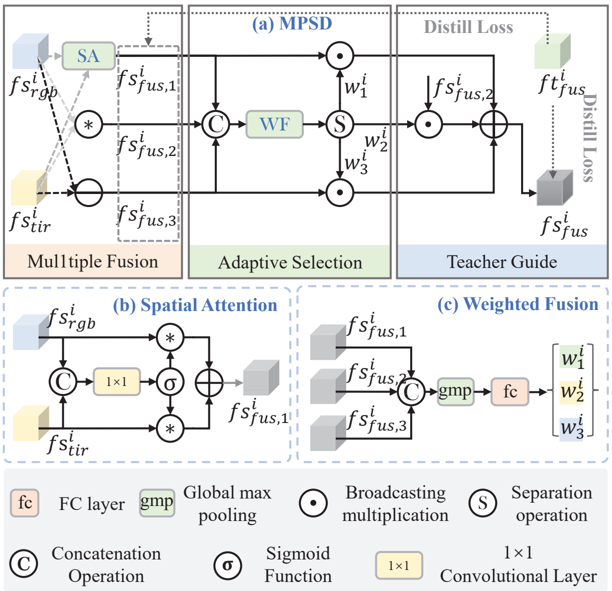

Figure 6. The proposed multi-path selection distillation module.

Here, $sa(*)$ denotes the spatial attention mechanism, which first utilizes a convolution layer of kernel size 1×1 and a softmax layer to get a two-channel weight map. The two-channel weight map is then split into two reliability weight maps for selecting the RGB features and TIR features, respectively. Mathematically, the self-attention mechanism is expressed by:

$$\begin{aligned} w^i_{rgb}, w^i_t &= \sigma(conv(cat(fs^i_{rgb}, fs^i_{tir}), \theta_1)), \\ fs^i_{fus,1} &= (fs^i_{rgb} \circledast w^i_{rgb}) \oplus (fs^i_{tir} \circledast w^i_{tir}), \end{aligned} \tag{5}$$

where $cat(*)$ denotes the concatenation operation and $conv(*, \theta_1)$ denotes a 1×1 convolutional layer with its parameters θ_1. $\sigma(*)$ denotes the sigmoid layer. The features $fs^i_{fus,1}$ mainly reflect the complementary information within multi-modal data. Features $fs^i_{fus,2}$ and $fs^i_{fus,3}$ reflect their interacted information and their differential information, respectively.

After that, $fs^i_{fus,1}$, $fs^i_{fus,2}$ and $fs^i_{fus,3}$ are further combined together by a weighted fusion way, i.e,

$$\begin{aligned} w^i_1, w^i_2, w^i_3 &= softmax(fc(gmp(cat(fs^i_{fus,1}, fs^i_{fus,2}, fs^i_{fus,3})))), \\ fs^i_{fus} &= (fs^i_{fus,1} \odot w^i_1) \oplus (fs^i_{fus,2} \odot w^i_2) \oplus (fs^i_{fus,3} \odot w^i_3). \end{aligned} \tag{6}$$

where $gmp(*)$ and $fc(*)$ denote the global max pooling layer and the fully connected layers, respectively. $softmax(*)$ denote the softmax operation. The feature-wise weights w^i_1, w^i_2, w^i_3 reflect the importance of different fused features for the current scenario. $\odot$ denotes the broadcasting multiplication operation.

With the fused features $\{ft^i_{fus}|i = 3,4\}$ and $\{fs^i_{fus}|i = 3,4\}$ obtained by the teacher model and the student model, respectively, we calculate the fusion distillation loss L_{fus}

between the fused features:

$$L_{fus} = l\left(ft^3_{fus}, fs^3_{fus}\right) + l\left(ft^4_{fus}, fs^4_{fus}\right) \tag{7}$$

What's more, in order to enable the student model to adaptively select a fusion path that is more similar to the teacher model in different scenarios, we introduce an additional penalty L_p for the efficiency of knowledge transformation during training. More specifically, we first select the fusion type with the smallest difference between the initially fused features of the student model and the fused features of the teacher model by,

$$\begin{aligned} L^i_{fus,n} &= l\left(ft^i_{fus}, fs^i_{fus,n}\right), n = 1,2,3 \\ L^i_{fus,\lambda^i} &= min(L^i_{fus,1}, L^i_{fus,2}, L^i_{fus,3}), \end{aligned} \tag{8}$$

where λ^i =1, 2 or 3 denotes the selected type of initially fused features according to the fused feature difference between the teacher and student models.

After that, through the adaptive selection part in MPSD, the student model itself will also predict a type of initially fused features that is suitable for the current tracking scene, i.e.,

$$w^i_{\nu^i} = max(w^i_1, w^i_2, w^i_3), \tag{9}$$

where ν^i =1, 2 or 3 denotes the predicted type of initially fused features from the student model.

With $w^i_{\nu^i}$ and $w^i_{\lambda^i}$, we can use a penalty to help the student model choose a fusion path that is more suitable for the current scene under the guidance of the teacher model, i.e.,

$$L_p = \sum_{i=3}^{4} max(|L^i_{fus,\lambda^i} w^i_{\nu^i} - L^i_{fus,\lambda^i} w^i_{\lambda^i}|, 0). \tag{10}$$

By minimizing L_p, $w^i_{\nu^i}$ and $w^i_{\lambda^i}$ will tend to be consistent, which can enable the student model to adaptively select the fusion path according to the teacher model to improve the exploration ability of complementary information.

On top of that, the overall distill loss in the multi-modal fusion stage can be obtained by:

$$L_{MPSD} = L_{fus} + L_p. \tag{11}$$

3.5. Hard-focused Response Distillation

To alleviate the data imbalance problem, we propose the Hard-focused Response Distillation (HFRD) module to instruct the student to focus on distinguishing targets from hard negative samples.

First, we obtain the response map $R_t \in \mathbb{R}^{H\times W}$ from the teacher model. Then, in order to prevent the teacher model from failing to have high responses in the target area within some scenes, we use the Gaussian-shaped mask $R_g \in \mathbb{R}^{H\times W}$ constructed by ground-truth bounding box as in [1] to correct the response map of the teacher model R_t as follows:

$$R_c(i,j) = \begin{cases} R_t(i,j) + R_g(i,j), \ if(R_t(i,j) + R_g(i,j)) < 1, \\ R_g(i,j), \ if(R_t(i,j) + R_g(i,j)) \geqq 1. \end{cases} \tag{12}$$

where i, j are the the horizontal and vertical coordinates of the response map, respectively. The corrected mask $R_c \in \mathbb{R}^{H\times W}$ has higher response values not only on the positive samples but also on the hard negative samples.

In the training process of the student model, with the assistance of the corrected mask R_c from the teacher model, the student model can focus more on distinguishing target from hard negative samples by a proposed Hard-focused Response Distillation loss L_{HFRD} to alleviate the data imbalance problem:

$$L_{HFRD} = r(R_s \circledast R_c, R_g), \tag{13}$$

where $r(*)$ denotes the L_2 loss function [5].

3.6. Overall loss

The overall distillation loss $L_{distill}$ is the sum of L_{SCFD}, L_{MPSD} and L_{HFRD}. We train the student model with the total loss as follows:

$$L_{distill} = \alpha(L_{SCFD} + L_{MPSD}) + \beta L_{HFRD} + L_{original}, \tag{14}$$

where α and β are hyper-parameters to balance the distillation loss. $L_{original}$ is the original loss for tracking as in [35]. The distillation loss L_{SCFD} and L_{MPSD} are just calculated on feature maps, which can be easily applied to different tackers or other multi-modal vision tasks.

4. Experiments

Our tracking approach is implemented in Python based on PyTorch. For inference, we test our tracker on a single Nvidia RTX 1080Ti GPU.

4.1. Implementation details

Training Details. We adopt the training dataset in LasHeR [14], which contains 979 pairs of RGB-T videos, to train the teacher model and student model, respectively. The proposed CMD framework includes two training stages. In the first stage, we train the teacher model as in MFNet [35] and fix its weights after training. In the second stage, the optimization of the student model is jointly supervised by the original tracking loss $L_{original}$ as well as the knowledge transfer loss L_{SCFD}, L_{MPSD} and L_{HFRD}. α and β in Eq. 14 are experimentally set to 0.001 and 100, respectively.

Online Tracking. In the tracking phase, our method is similar to DiMP [1]. We split tracking into classification and regression subtasks. For classification subtask, we employ data augmentation [1] on the first frame to construct an initial set, which contains 15 initial training samples for initial classification model training. Then the initial classification model is optimized using the augmented training set during tracking. For regression subtask, the same settings as those in [1] are employed.

Table 1. Ablation study of different components.

Student	SCFD	MPSD	HFRD	PR/SR	Params(M)	FPS
✓				78.1/56.0	19.8	31
✓	✓			80.3/57.4	19.8	31
✓	✓	✓		82.2/58.3	19.9	30
✓	✓	✓	✓	82.4/58.4	19.9	30

Table 2. Ablation study of the proposed SCFD module.

Student	AFD	SED	CED	SCFD	PR/SR	Params(M)	FPS
✓					78.1/56.0	19.8	31
✓	✓				78.7/56.6	19.8	31
✓		✓			79.2/56.5	19.8	31
✓			✓		79.6/56.6	19.8	31
✓				✓	80.3/57.4	19.8	31

Table 3. Ablation study of the proposed MPSD module.

Student+SCFD	SAF	CAF	TF	MPSD	PR/SR	Params(M)	FPS
✓					80.3/57.4	19.8	31
✓	✓				81.9/58.0	19.9	30
✓		✓			81.1/57.6	19.9	30
✓			✓		82.6/58.6	25.9	23
✓				✓	82.2/58.3	19.9	30

4.2. Evaluation datasets and metrics

We evaluate our method on three large-scale benchmark datasets, i.e., GTOT [10], RGBT234 [11] and LasHeR [14]. GTOT is the first standard dataset for RGB-T tracking. It contains 50 RGB-T video sequences annotated with seven challenging attributes. RGBT234 [11] contains 234 pairs of RGB-T videos and 12 annotated attributes. LasHeR is currently the largest RGB-T tracking dataset, which consists of 1244 RGB-T videos with more than 730K frame pairs in total. Among them, 245 videos are used as the testing set, and 979 videos are used as the training set. As in [17], we utilize two widely used metrics, i.e., precision rate (PR) and success rate (SR), to evaluate the tracking performance on GTOT and RGBT234. As in [14], we adopt precision rate (PR), normalized precision rate (NPR) and success rate (SR) to evaluate different trackers on LasHeR.

4.3. Ablation Experiments and Analyses

We conduct some ablation studies on RGBT234 [11] to discuss the impacts of different components in our CMD framework.

Ablation experiments for each module. To investigate the impact of each component in our proposed CMD, several versions of our proposed method are provided for comparisons. Specifically, 'Student' denotes the model that without any knowledge transformation. Here, the proposed SCFD, MPSD and HFRD are employed in the unimodal feature extraction stage, multi-modal feature fusion stage and target state estimation stage, respectively. The quantitative results of these models are shown in Table 1. It can be seen that SCFD, MPSD and HFRD can all improve the performance of the student model. This verifies that each proposed component in CMD can effectively inherit the knowledge learnt from a powerful teacher models to a student model without obvious loss.

Effectiveness of the proposed SCFD module. To further verify the effectiveness of the proposed SCFD module, several variants are also compared with our proposed SCFD module. Here, in the unimodal feature extraction stage, 'AFD', 'SED' and 'CED' denote initially fusing the unimodal features of each layer in the teacher model by using the simple element-wise addition operation, the designed SEM module and the designed CEM module, respectively, and such initially fused features are then employed to guide the student model for single-stream structure learning. As well, ABF modules [3] are employed in all of these variants. As shown in Table 2, the proposed SCFD can better exploit the modality-common and modality-specific information from the teacher model.

Effectiveness of the proposed MPSD module. As shown in Table 3, several versions of our proposed MPSD module are also conducted to verify its effectiveness. 'SAF' denotes a spatial-wise attention based fusion module. 'CAF' denotes a channel-wise attention based module. 'TF' denotes adopting the same fusion strategy as that in the teacher model [35]. In particular, each layer adopts the same fusion strategy in the student model. It can be seen that the exploitation of the multi-path fusion strategy can well improve the performance of the student model. In addition, the performance gap between 'Student-MPSD' and 'Student-TF' is much smaller, which indicates that our proposed MPSD module can better mimic the fused features in the teacher model to compensate for the performance penalty from simple fusion operations.

Teacher-Student knowledge distillation experiments. Table 4 shows the performance of using some other knowledge distillation methods in the feature extraction and feature fusion stages for comparisons, including KD [8], FitNets [21], ReviewKD [3] and MD [29]. It is observed from Table 4 that the proposed distillation strategy performs the best. Due to the absence of cross-modal interactions, these existing knowledge distillation methods usually achieve some modest performance gains. In addition, we notice that the student model with a single-stream feature extractor performs obviously well than the student model with a two-stream feature extractor after knowledge distillation. This may be due to the fact that the single-stream network can narrow the modality difference to a certain extent and better acquire the knowledge from the teacher model.

4.4. Comparison with the state-of-the-art

To evaluate the superiority of our proposed method, we compare our method with some existing state-of-the-art RGB-T trackers, including MANet [17], DAFNet [6], DAPNet [39], TODA [28], MACNet [30], CAT [12], CEDiMP

Table 4. Ablation study of different knowledge distillation experiments.

Method	KD [8]	FitNets [21]	ReviewKD [3]	ReviewKD [3]	MD [29]	Ours
Backbone	2 Res18	2 Res18	2 Res18	1 Res18	1 Res18	1 Res18
PR/SR	78.4/56.2	79.3/56.0	79.9/56.5	80.5/56.3	80.0/55.4	82.4/58.4
FPS	27	27	27	30	30	30
Params	31.6	31.6	31.6	19.9	19.9	19.9

Table 5. Quantitative comparisons of our method with some state-of-the-arts methods on benchmark datasets. Higher values indicate better performance.

Methods	Year	RGBT234 [11]		GTOT [10]		LasHeR [14]			FPS	Params
		PR ↑	SR ↑	PR ↑	SR↑	PR ↑	NPR ↑	SR ↑		
DAPNet [39]	2019	76.6	53.7	88.2	70.7	43.1	38.4	31.4	1	-
MANet [17]	2019	77.7	53.9	89.4	72.4	45.7	40.8	33.0	1	7.28M
DAFNet [6]	2019	79.6	54.4	89.1	71.6	44.9	39.0	31.1	14	5.50M
mfDiMP [31]	2019	78.6	55.5	83.6	69.7	44.7	39.5	34.4	22	175.82M
TODA [28]	2019	78.7	54.5	84.3	67.7	-	-	-	1	-
MACNet [30]	2020	79.0	55.4	88.0	71.4	48.3	42.3	35.2	1	14.86M
CAT [12]	2020	80.4	56.1	88.9	71.7	45.1	39.8	31.7	-	-
FANet [40]	2021	78.7	55.3	89.1	72.8	44.2	38.4	30.9	12	38.44M
SiamCDA [36]	2021	76.0	56.9	87.7	73.2	-	-	-	24	107.90M
JMMAC [32]	2021	79.0	57.3	90.2	73.2	-	-	-	-	-
MANet++ [33]	2021	80.0	55.4	88.2	70.7	46.7	40.8	31.7	15	7.38M
ADRNet [33]	2021	80.9	57.1	90.4	73.9	-	-	-	15	68.50M
CBPNet [27]	2022	79.4	54.1	88.5	71.6	-	-	-	3	-
TFNet [41]	2022	80.6	56.0	88.6	72.9	-	-	-	-	-
MFGNet [25]	2022	78.3	53.5	88.9	70.7	-	-	-	3	8.09M
M5LNet [23]	2022	79.5	54.2	89.6	71.0	-	-	-	9	-
HMFT [34]	2022	78.8	56.8	91.2	74.9	-	-	-	-	127.84M
APFNet [26]	2022	82.7	57.9	90.5	73.9	50.0	-	36.2	-	15.01M
MANet* [17]	2019	78.6	55.5	90.0	72.5	-	-	-	1	7.28M
DAFNet* [6]	2019	80.0	54.9	86.0	70.0	48.0	42.8	34.5	14	5.50M
mfDiMP* [31]	2019	82.4	58.3	87.7	73.1	58.3	54.2	45.6	22	175.82M
FANet* [40]	2021	79.4	53.9	90.1	72.1	48.2	42.5	34.3	12	38.44M
Teacher [35]	2022	84.4	60.1	90.7	73.5	59.7	55.4	46.7	18	81.01M
Student-Origin	2022	78.1	56.0	88.5	72.3	55.4	50.3	42.3	30	19.90M
Student-Distill	2022	82.4	58.4	89.2	73.4	59.0	54.6	46.4	30	19.90M

[38], SiamCDA [36], mfDiMP [31], FANet [40], CBPNet [27], MANet++ [33], JMMAC [32], ADRNet [33], TFNet [41], MFGNet [25], M5LNet [23], HMFT [34], APFNet [26] and MFNet [35], on three challenging datasets. Considering that existing methods usually employ different training datasets, we use the LasHeR training set to retrain some of these algorithms for fair comparisons, including FANet*, DAFNet*, MANet* and mfDiMP*.

On RGBT234. From Table 5, we observe that, in addition to the teacher model, our method achieves the best results with 82.4%/58.4% in PR/SR. In particular, our tracker achieves 3.0%/4.5%, 3.8%/2.9% and 2.4%/3.6% improvements against FANet* [40], MANet* [17] and DAFNet* [6] in PR/SR, respectively. We should note that some MDNet [18] based trackers [6, 17, 33] employ merely three convolutional layers for unimodal feature extraction, thus have fewer parameters in the unimodal feature extraction stage with a two-stream structure (e.g., 3.6M in [30, 40]). However, in the multi-modal feature fusion stage, these methods' parameters are significantly higher than those of the proposed MPSD module. Meanwhile, due to the complex updating strategy in MDNet, the above methods cannot meet the needs of real-time operation.

On GTOT. From Table 5, we can see that our method obtains competitive performance on GTOT dataset [10] with 73.4% and 89.2% in success and precision scores, respectively. Compared with the teacher model, our student model achieves comparable performance but with a 75% reduction in parameter sizes. Compared with the original student model, our algorithm achieves 1.1% improvements in success and 0.7% improvements in precision.

On LasHeR. LasHeR [14] is captured from a number of scenes and categories and is highly diverse. A tracker retrained on this dataset usually achieves some improvements. From Table 5, we can also see that, in addition to the teacher model, our tracker still performs the best in terms of all the three metrics with significant performance superiorities on LasHeR. In particular, our tracker achieves 11.0%/11.9% and 10.8%/12.1% improvements against DAFNet* [6] and FANet* [40], which are based on MDNet [18]. Compared with mfDiMP* [31], which is based on DiMP [1] and employs two ResNet50 [7] for feature extraction, our PR/SR is 0.7%/0.8% higher than it. This demonstrates that our proposed method can effectively reduce the performance loss caused by parameter reduction.

5. Conclusion

In this paper, a novel teacher-student knowledge distillation training framework is proposed to reduce the performance gap between a powerful teacher model and a compact student model. Specifically, this framework distills the knowledge from a deep two-stream network with complex multi-modal feature fusion modules to a single-stream network with efficient feature fusion modules. By virtue of the proposed SCFD module, the modality-common information as well as the modality-specific information can be transformed from a two-stream network to a single-stream network in the unimodal feature extraction stage, thus enhancing the representations of unimodal features. Besides, by employing the proposed MPSD module, the student model can adaptively combine multiple fused features generated by various simple fusion strategies to explore complementary information from multi-modal data more thoroughly. In addition, an HFRD module is proposed to improve the student model's discriminative ability against the distractors by alleviating the problem of data imbalance in the target state estimation stage. Experimental results show that our approach helps a student model achieves the state-of-the-art performance while reducing the number of parameters and computational complexity dramatically.

Limitation: The current method dedicated to reducing computational complexity at the stages of unimodal feature extraction and multi-modal feature fusion, but it paid $ZERO$ effort to improve the efficiency of target stage estimation, which is our future work.

Acknowledgement: This work is supported by the National Natural Science Foundation of China under Grant No. 61773301. It is also supported by the Shaanxi Innovation Team Project under Grant No.2018TD-012.

References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6182–6191, 2019. 3, 4, 6, 8

[2] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017. 3

[3] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2021. 5, 7, 8

[4] Ciarán Ó Conaire, Noel E O'Connor, and Alan Smeaton. Thermo-visual feature fusion for object tracking using multiple spatiogram trackers. *Machine Vision and Applications*, 19(5-6):483–494, 2008. 2

[5] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: Accurate tracking by overlap maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4660–4669, 2019. 4, 6

[6] Yuan Gao, Chenglong Li, Yabin Zhu, Jin Tang, Tao He, and Futian Wang. Deep adaptive fusion network for high performance rgbt tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019. 2, 7, 8

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 4, 8

[8] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 3, 7, 8

[9] Nianchang Huang, Qiang Jiao, Qiang Zhang, and Jungong Han. Middle-level feature fusion for lightweight rgb-d salient object detection. *IEEE Transactions on Image Processing*, 2022. 2

[10] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25(12):5743–5756, 2016. 2, 7, 8

[11] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. 1, 2, 7, 8

[12] Chenglong Li, Lei Liu, Andong Lu, Qing Ji, and Jin Tang. Challenge-aware rgbt tracking. In *European Conference on Computer Vision*, pages 222–237. Springer, 2020. 7, 8

[13] Chenglong Li, Xiang Sun, Xiao Wang, Lei Zhang, and Jin Tang. Grayscale-thermal object tracking via multitask laplacian sparse representation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(4):673–681, 2017. 2

[14] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, and Jin Tang. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *arXiv preprint arXiv:2104.13202*, 2021. 2, 6, 7, 8

[15] Tianshan Liu, Kin-Man Lam, Rui Zhao, and Guoping Qiu. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):315–329, 2022. 2, 3

[16] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 3

[17] Cheng Long Li, Andong Lu, Ai Hua Zheng, Zhengzheng Tu, and Jin Tang. Multi-adapter RGBT tracking. In *Proceedings of the IEEE Conference on Computer Vision Workshops*, 2019. 2, 7, 8

[18] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4293–4302, 2016. 2, 8

[19] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9060–9069, 2020. 2, 3

[20] Guangyu Ren, Yinxiao Yu, Hengyan Liu, and Tania Stathaki. Dynamic knowledge distillation with noise elimination for rgb-d salient object detection. *Sensors*, 22(16):6188, 2022. 2, 3

[21] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 3, 5, 7, 8

[22] Jianbing Shen, Yuanpei Liu, Xingping Dong, Xiankai Lu, Fahad Shahbaz Khan, and Steven CH Hoi. Distilled siamese networks for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[23] Zhengzheng Tu, Chun Lin, Wei Zhao, Chenglong Li, and Jin Tang. M5l: Multi-modal multi-margin metric learning for rgbt tracking. *IEEE Transactions on Image Processing*, 31:85–98, 2022. 8

[24] Ning Wang, Wengang Zhou, Yibing Song, Chao Ma, and Houqiang Li. Real-time correlation tracking via joint model compression and transfer. *IEEE Transactions on Image Processing*, 29:6123–6135, 2020. 3

[25] Xiao Wang, Xiujun Shu, Shilliang Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Mfgnet: Dynamic modality-aware filter generation for rgb-t tracking. *IEEE Transactions on Multimedia*, pages 1–1, 2022. 8

[26] Yun Xiao, Mengmeng Yang, Chenglong Li, Lei Liu, and Jin Tang. Attribute-based progressive fusion network for rgbt tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 8

[27] Q. Xu, Y. Mei, J. Liu, and C. Li. Multimodal cross-layer bilinear pooling for rgbt tracking. *IEEE Transactions on Multimedia*, pages 1–1, 2021. 8

[28] R. Yang, Y. Zhu, X. Wang, C. Li, and J. Tang. Learning target-oriented dual attention for robust rgb-t tracking. In *Proceedings of IEEE International Conference on Image Processing*, pages 3975–3979, 2019. 7, 8

[29] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Low-cost multispectral scene analysis with modality distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 803–812, 2022. 2, 3, 7, 8

[30] Hui Zhang, Lei Zhang, Li Zhuo, and Jing Zhang. Object tracking in RGB-T videos using modal-aware attention network and competitive learning. *Sensors*, 20(2):393, 2020. 2, 7, 8

[31] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan. Multi-modal fusion for end-to-end RGB-T tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019. 2, 8

[32] Pengyu Zhang, Jie Zhao, Chunjuan Bo, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Jointly modeling motion and appearance cues for robust rgb-t tracking. *IEEE Transactions on Image Processing*, 30:3335–3347, 2021. 8

[33] Pengyu Zhang, Jie Zhao, Chunjuan Bo, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Jointly modeling motion and appearance cues for robust rgb-t tracking. *IEEE Transactions on Image Processing*, 30:3335–3347, 2021. 8

[34] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8886–8895, June 2022. 8

[35] Qiang Zhang, Xueru Liu, and Tianlu Zhang. Rgb-t tracking by modality difference reduction and feature re-selection. *Image and Vision Computing*, 127:104547, 2022. 1, 2, 3, 4, 6, 7, 8

[36] Tianlu Zhang, Xueru Liu, Qiang Zhang, and Jungong Han. SiamCDA: Complementarity-and distractor-aware RGB-T tracking based on Siamese network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 3, 8

[37] Xingchen Zhang, Ping Ye, Shengyun Peng, Jun Liu, Ke Gong, and Gang Xiao. SiamFT: An RGB-Infrared fusion tracking method via fully convolutional Siamese networks. *IEEE Access*, 7:122122–122133, 2019. 3

[38] Long Zhao, Meng Zhu, Honge Ren, and Lingjixuan Xue. Channel exchanging for rgb-t tracking. *Sensors*, 21(17):5800, 2021. 8

[39] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. Dense feature aggregation and pruning for RGBT tracking. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 465–472, 2019. 2, 7, 8

[40] Y. Zhu, C. Li, J. Tang, and B. Luo. Quality-aware feature aggregation network for robust rgbt tracking. *IEEE Transactions on Intelligent Vehicles*, 6(1):121–130, 2021. 8

[41] Yabin Zhu, Chenglong Li, Jin Tang, Bin Luo, and Liang Wang. Rgbt tracking by trident fusion network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):579–592, 2021. 8